

# 先端人工知能論I

東京大学 大学院情報理工学系研究科

牛久 祥孝

# 第9回～第11回の内容と目標

- **基礎：**  
系列データの理解/生成に用いられる  
Recurrent Neural Networks (RNNs) の理解
- **応用：**  
自然言語処理や画像 + 言語融合分野の理解

Chap. 9 RNN

Chap. 10 Long Short-Term Memory (LSTM)  
自然言語処理 (NLP)

**Chap. 11 NLPと画像理解**

# 第11回の内容と目標

## 1. 座学 NLPと画像理解

- LSTMとword2vec (SGNS)の復習
- ニューラル機械翻訳とアテンション
- 画像キャプション生成

## 2. 演習 アテンションを用いた機械翻訳モデルの実装

## 3. 演習 アテンションを用いたキャプション生成モデルの実装

## 4. 座学 NLPと画像理解の発展

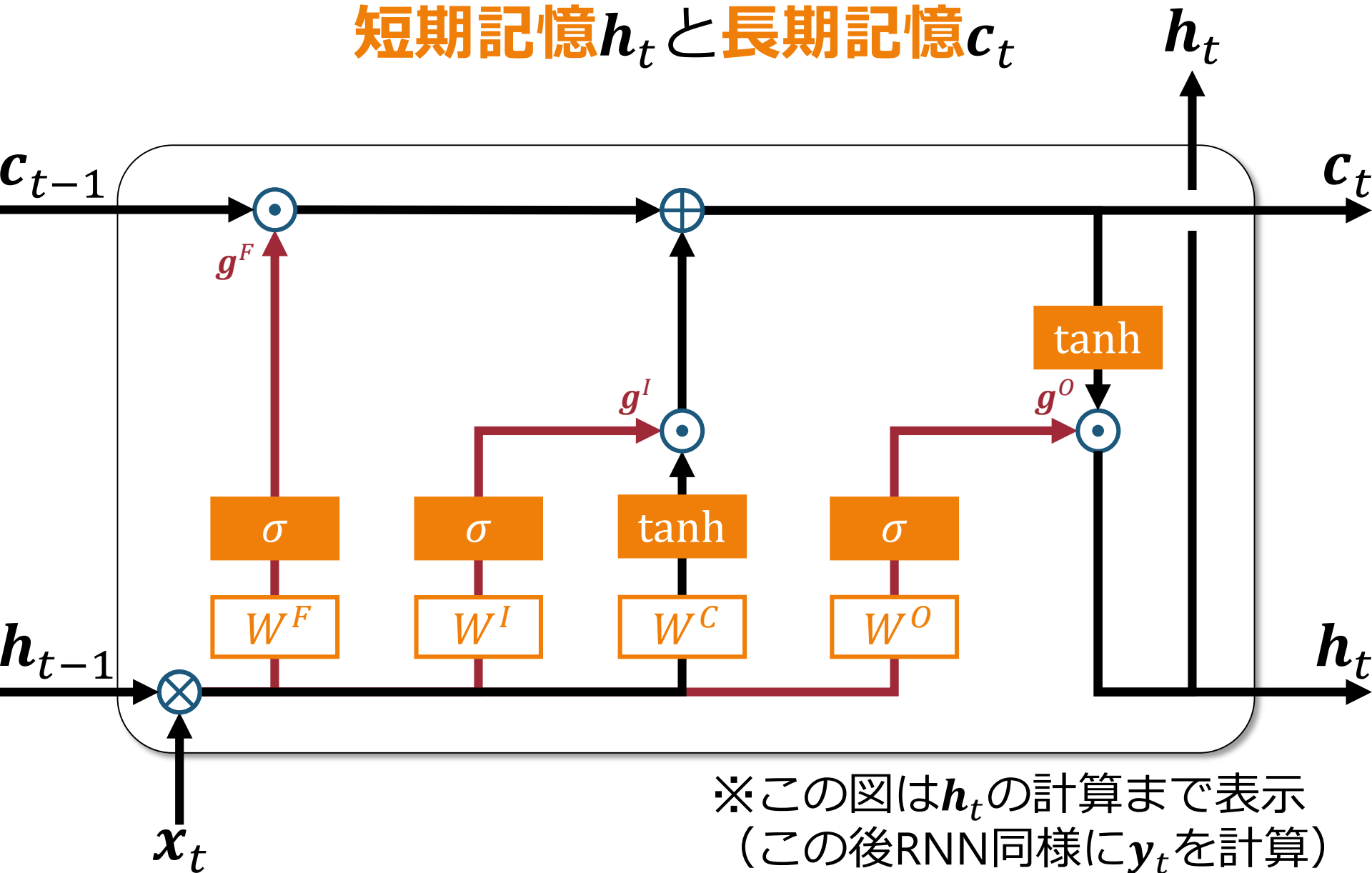
- 画像キャプション生成やその他の課題へ

**NLPと画像理解**

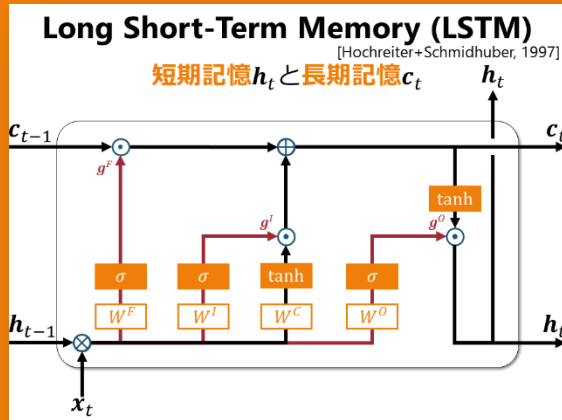
# Long Short-Term Memory (LSTM)

[Hochreiter+Schmidhuber, 1997]

短期記憶 $h_t$ と長期記憶 $c_t$



# LSTMは一見複雑



狂気の沙汰としか思えない...

## 1ステップずつ理解 すれば怖くない！

### 1. 長期記憶の更新

1. 忘却
2. 追加

### 2. 短期記憶の更新

## 表の見方

**tanh** 活性化関数

**$W$**  線形変換



ベクトルの直列



ベクトルの要素ごとの和



ベクトルの要素ごとの積



# もっとも単純な方法

- **One-hot ベクトル**

- 1-of-K ベクトルともいう
- 単語の種類の数（語彙数）と同じ次元
- ある単語が対応する次元だけ1、他は0

- **例：“Language”, “Natural”, “Processing”  
という言葉しかない世界では...**

Natural Language Processing

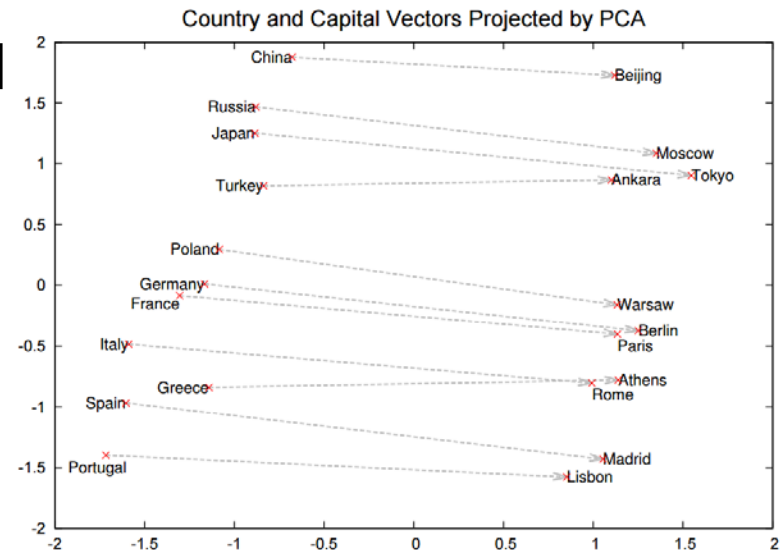
→ “Natural”, “Language”, “Processing”

→  $\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$  ← Language の次元  
← Natural の次元  
← Processing の次元

# 分散表現

数百次元と低次元で、単語間の類似度が埋め込まれた空間内での、各単語のベクトル表現

- 深層学習による分散表現
- 深層学習によらない分散表現
  - “Word2vec” [Mikolov+, NIPS 2014]  
(Skip-gram with Negative Sampling; SGNS)
  - GloVe [Pennington+, EMNLP 2015]





# Neural Machine Translation (NMT)

## EncoderとDecoderからなる系列変換モデルを利用した機械翻訳手法

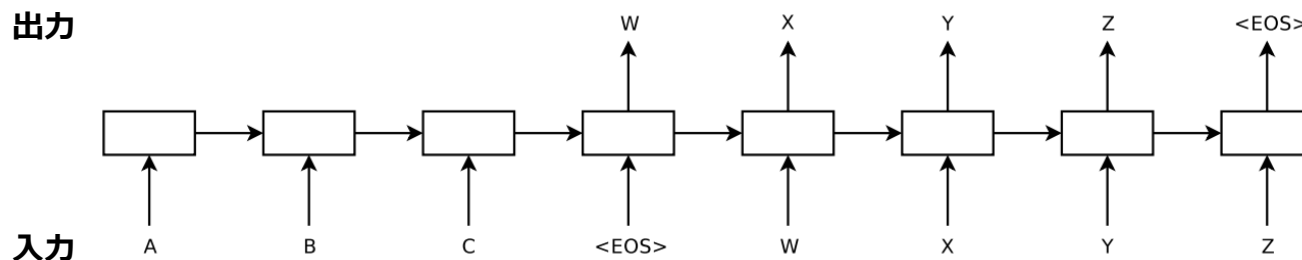
- 2011 音声認識で深層学習がSOTA
- 2012 画像認識で深層学習がSOTA
- 2014 sequence2sequenceモデル提案**
- 2015 機械翻訳で深層学習がSOTA**  
**→NMTの隆盛**

※SOTA = state-of-the-art

# sequence2sequence

[Sutskever+, NIPS 2014]

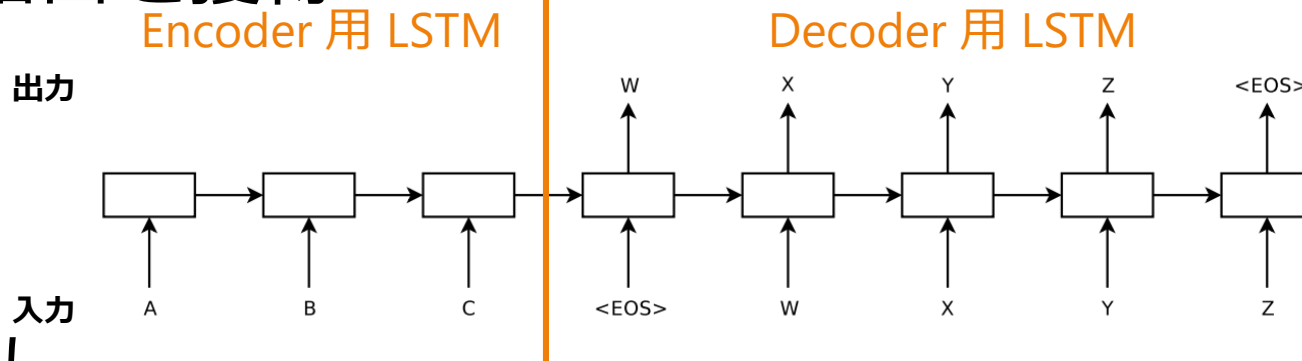
- Encoderに単語を一つずつ入力して隠れ変数 $h$ を計算する
- 隠れ変数を $h$ としたDecoderに<EOS>を入力して1単語目を獲得
- <EOS>が出るまで、 $n - 1$ 番目の単語を入力して $n$ 単語目を獲得



# sequence2sequence

[Sutskever+, NIPS 2014]

- Encoderに単語を一つずつ入力して隠れ変数 $h$ を計算する
- 隠れ変数を $h$ としたDecoderに<EOS>を入力して1単語目を獲得
- <EOS>が出るまで、 $n - 1$ 番目の単語を入力して $n$ 単語目を獲得



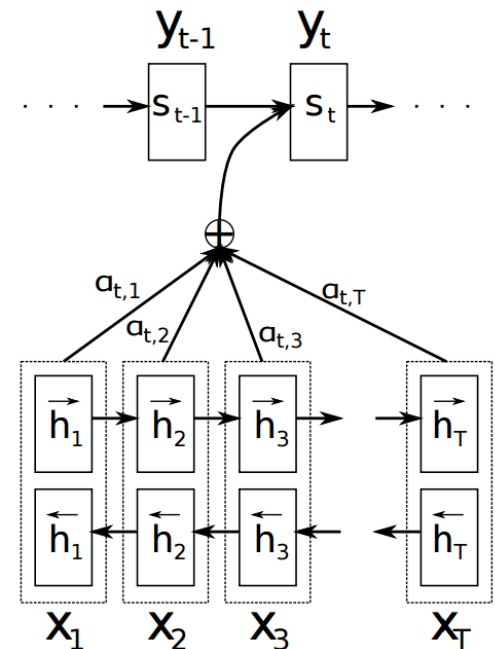
- ただし...
  - 2つの異なるLSTMが同じように並んでいるので注意
  - 入力文の単語を逆から入れる  
(順方向で入れると、LSTMでも文頭を忘れてしまう)

# アテンションを用いた機械翻訳

- **sequence2sequence**は...
  - 一旦EncodeしたらあとはDecoderに任せる
  - 長い文は扱えないのでは

- **アテンションの利用** [Bahdanau+, ICLR 2015]

- アテンションとは...
  - 「 $t$ 番目の単語を出力する時に、  
入力文のどこを翻訳すればよいか」
  - 入力文の単語数 $T$ と同じ数の  
ベクトル  $a_t$  を計算
- 隠れ変数  $h_t$  の重みづけ和を計算  
→ LSTMへ入力



# sequence2sequence with attention

[Luong+, EMNLP 2015]

- 2つのアテンションモデルを提案（後述）

- 局所的アテンション（発展的、より高性能）
- 大域的アテンション（本講義の演習で採用）

- Input-feeding

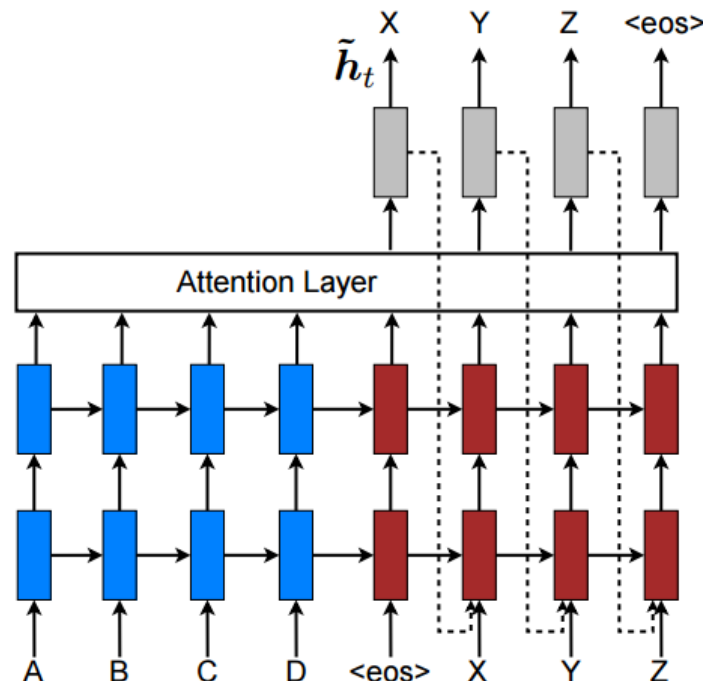
- 一個前の単語だけではなく隠れ変数 $h$ も入力

- 定性的には：

直前にアテンションを  
あてた位置を知らせる

- ソースコード公開済み

- seq2seq-attn
- OpenNMT



# sequence2sequence with attention

[Luong+, EMNLP 2015]

## • 大域的アテンション

- $t$ 番目の隠れ変数を仮決め  $\mathbf{h}_t = \text{LSTM}(\mathbf{h}_{t-1})$
- 入力文のどこにアテンションをあてるかを計算

$$\mathbf{a}_t(s) = \text{align}(\mathbf{h}_t, \bar{\mathbf{h}}_s) = \frac{\exp(\mathbf{h}_t^\top \mathbf{W}_a \bar{\mathbf{h}}_s)}{\sum_{s'} \exp(\mathbf{h}_t^\top \mathbf{W}_a \bar{\mathbf{h}}_{s'})}$$

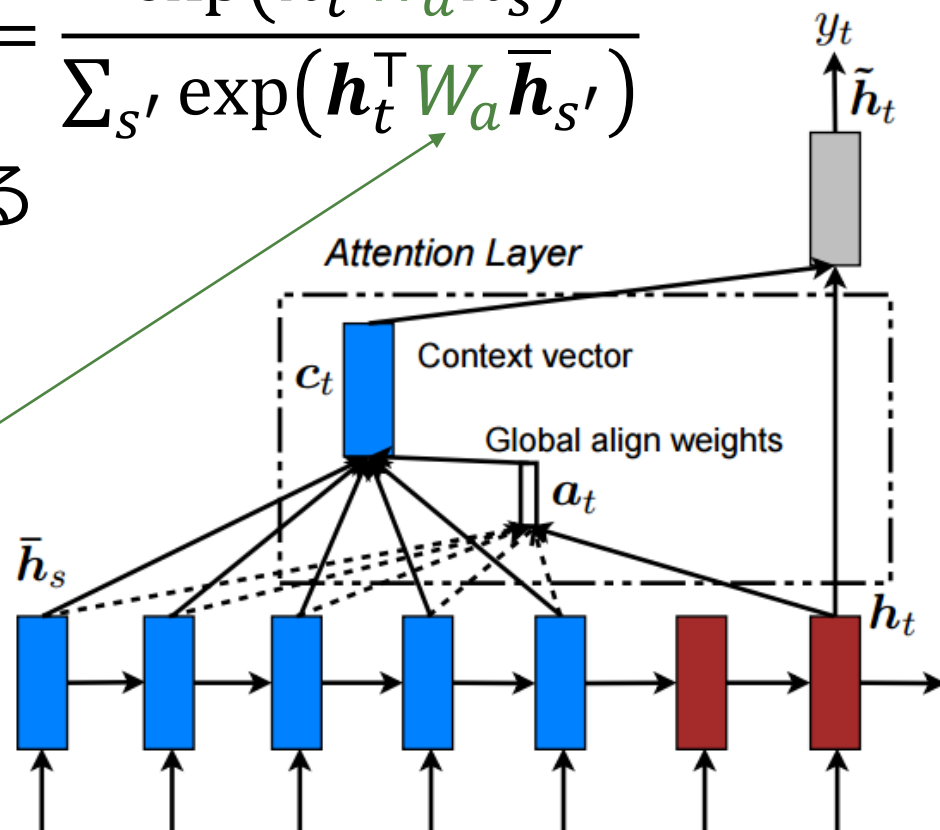
- コンテキストを求める

$$\mathbf{c}_t = \sum_s \mathbf{a}_t(s) \bar{\mathbf{h}}_s$$

- 最終的な隠れ変数

$$\tilde{\mathbf{h}}_t = \tanh(\mathbf{W}_c [\mathbf{c}_t; \mathbf{h}_t])$$

学習すべきパラメータ



# sequence2sequence with attention

[Luong+, EMNLP 2015]

## • 局所的アテンション

- 隠れ変数  $h_t$  から、アテンションをあてる位置  $p_t$  を計算

$$p_t = S \cdot \text{sigmoid} \left( v_p^T \tanh(W_p h_t) \right)$$

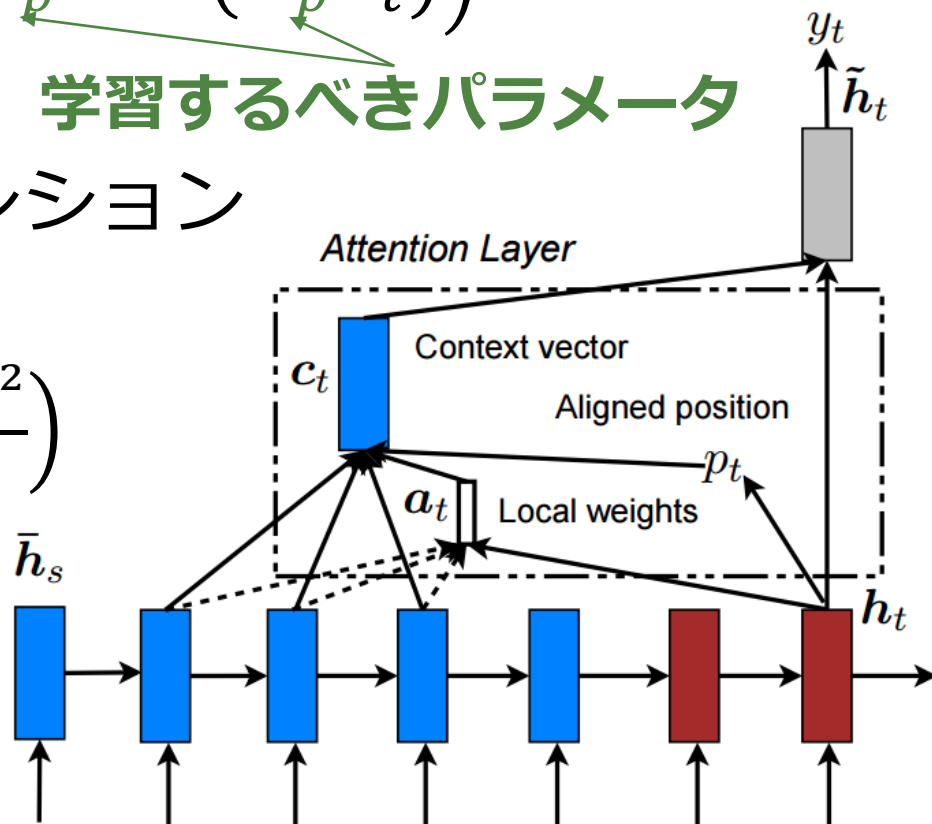
文長

学習すべきパラメータ

- $p_t$  を中心としたアテンション

$$a_t(s) = \text{align}(h_t, \bar{h}_s) \cdot \exp \left( -\frac{(s-p_t)^2}{2\sigma^2} \right)$$

- 後の計算は同様

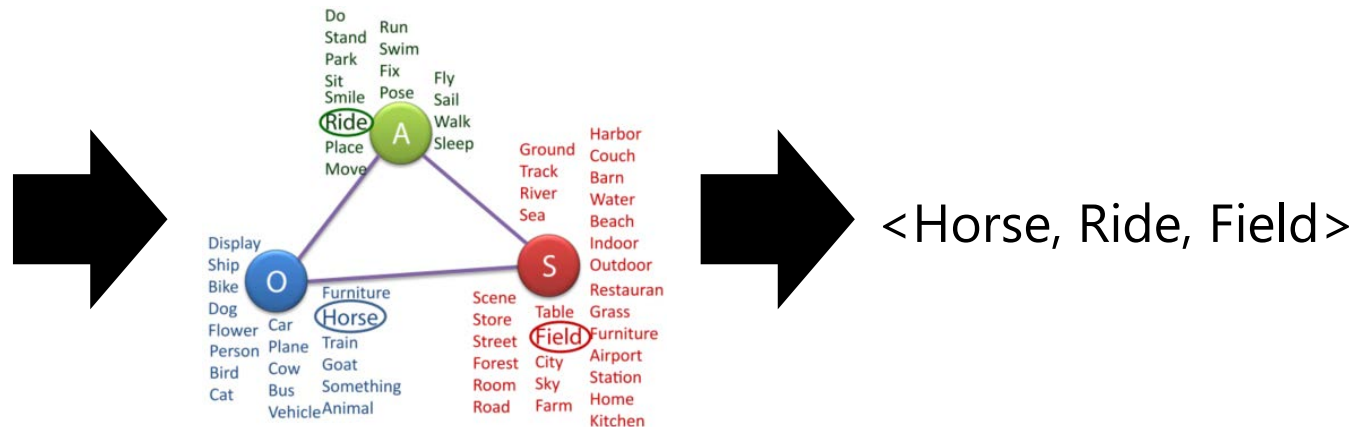


# Every picture tells a story [Farhadi+, ECCV 2010]

データセット：

画像 + <object, action, scene> + キャプション





## 1. 画像の<object, action, scene>をMRFで推定



## 2. <object, action, scene>が同じキャプションを検索して利用



# Every picture tells a story [Farhadi+, ECCV 2010]

	<p>(pet, sleep, ground)          (dog, sleep, ground)          (animal, sleep, ground)          (animal, stand, ground)          (goat, stand, ground)</p>	<p>see something unexpected.          Cow in the grassfield.          Beautiful scenery surrounds a fluffly sheep.          Dog hearding sheep in open terrain.          Cattle feeding at a trough.</p>
	<p>(furniture, place, furniture)          (furniture, place, room)          (furniture, place, home)          (bottle, place, table)          (display, place, table)</p>	<p>Refrigerator almost empty.          Foods and utensils.          Eatables in the refrigerator.          The inside of a refrigerator apples, cottage cheese, tupperwares and lunch bags.          Squash apenny white store with a hand statue, picnic tables in front of the building.</p>
	<p>(transportation, move, track)          (bike, ride, track)          (transportation, move, road)          (pet, sleep, ground)          (bike, ride, road)</p>	<p>A man stands next to a train on a cloudy day          A backpacker stands beside a green train          This is a picture of a man standing next to a green train          There are two men standing on a rocky beach, smiling at the camera.          This is a person laying down in the grass next to their bike in front of a strange white building.</p>
	<p>(display, place, table)          (furniture, place, furniture)          (furniture, place, furniture)          (bottle, place, table)          (furniture, place, home)</p>	<p>This is a lot of technology.          Somebody's screensaver of a pumpkin          A black laptop is connected to a black Dell monitor          This is a dual monitor setup          Old school Computer monitor with way to many stickers on it</p>

# 再利用？新規生成？

入力



データセット



A small white dog wearing a flannel warmer.



A small gray dog on a leash.



A black dog standing in grassy area.

- 再利用

- 新規生成

- テンプレート

- 主語 + 動詞の文を生成しよう

- 非テンプレート

# 再利用？新規生成？

入力



データセット



A small white dog wearing a flannel warmer.



A small gray dog on a leash.



A black dog standing in grassy area.

- **再利用**

- A small gray dog on a leash.

- **新規生成**

- テンプレート

- 主語 + 動詞の文を生成しよう

- 非テンプレート

# 再利用？新規生成？

入力



データセット



A small white dog wearing a flannel warmer.



A small gray dog on a leash.



A black dog standing in grassy area.

- **再利用**

- A small gray dog on a leash.

- **新規生成**

- テンプレート

- $\text{dog} + \text{stand} \Rightarrow \text{A dog stands.}$

- 非テンプレート

# 再利用？新規生成？

入力



データセット



A small white dog wearing a flannel warmer.



A small gray dog on a leash.



A black dog standing in grassy area.

- **再利用**

- A small gray dog on a leash.

- **新規生成**

- テンプレート

- $\text{dog} + \text{stand} \Rightarrow \text{A dog stands.}$

- 非テンプレート

- A small white dog standing on a leash.

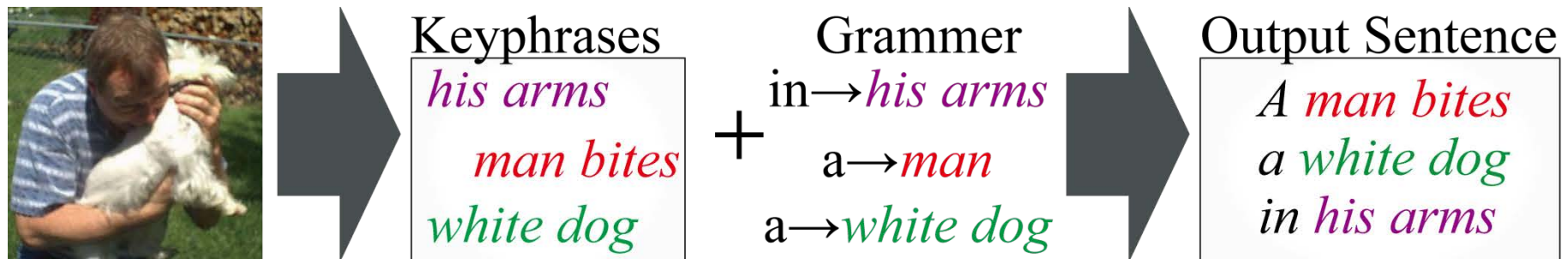


# マルチキーフレーズ推定アプローチ

[Ushiku+, ACM MM 2012]

当時の問題 = 使用候補であるフレーズの精度が悪い

**仮説: 画像の内容は少数の主要なフレーズで特定可能  
あとは文法モデルで繋げばよい!**



キーフレーズを独立なラベルとして扱うと...

**マルチキーフレーズの推定 = 一般画像認識**

文生成は[Ushiku+, ACM MM 2011]と同じ

Input Image

Keyphrases

Sentence



field EOS  
front of  
in front  
a black  
tracks EOS

A black and white  
cow in front  
of a man.



and  
sitting on  
a woman  
in front  
front of

Front of a  
woman in front  
of people  
sitting on.



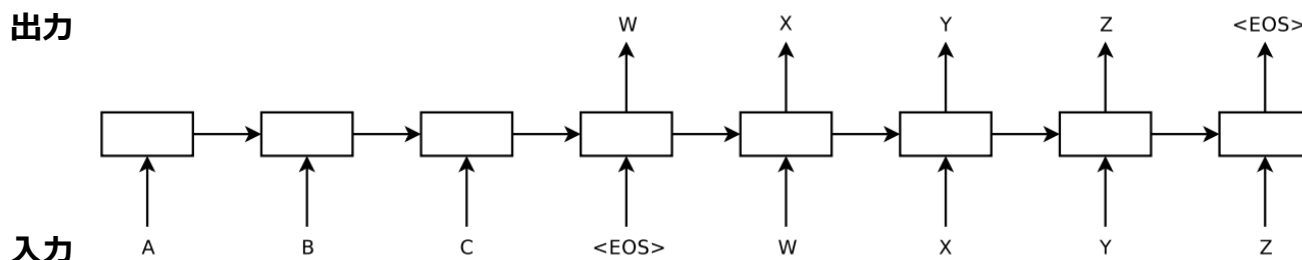
a brown  
water EOS  
field EOS  
brown horse  
a horse

Sandy field  
with a brown  
horse standing  
in a horse.

文の終わり

# Deep Learning の恩恵 (再掲)

- 深層学習による画像認識の精緻化 [Krizhevsky+, NIPS 2012]
- 機械翻訳でも深層学習が登場 [Sutskever+, NIPS 2014]
  - RNNで問題になっていた勾配の消失をLSTM [Hochreiter+Schmidhuber, 1997] で解決  
→文中の離れた単語間での関係を扱えるように



- LSTMを4層つなぎ、end-to-endで機械学習  
→state-of-the-art並み (英仏翻訳)

CNN/RNNなどの共通技術が台頭

➡ 画像認識や機械翻訳の参入障壁が低下

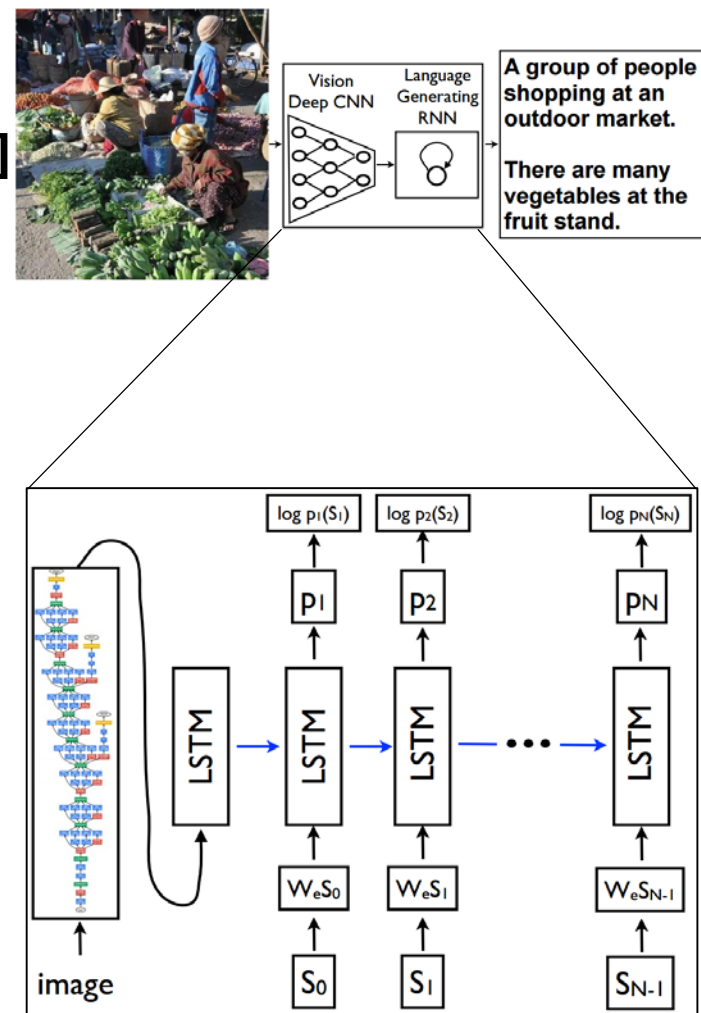


# Google NIC [Vinyals+, CVPR 2015]

## Googleで開発された

- **GoogLeNet** [Szegedy+, CVPR 2015]
  - **LSTM** [Sutskever+, NIPS 2014]
- を直列させて文生成する。

画像 $I$ への文（単語列） $S_0 \dots S_N$ は  
 $S_0$ : スタートを意味する単語  
 $S_1 = \text{LSTM}(\text{CNN}(I))$   
 $S_t = \text{LSTM}(S_{t-1}), t = 2 \dots N - 1$   
 $S_N$ : ストップを意味する単語



# 生成された説明文の例

**A person on a beach  
flying a kite.**

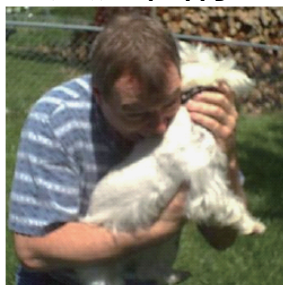


**A black and white photo of  
a train on a train track.**



# [Ushiku+, ACM MM 2012]と比べると

入力画像



Keyphrases  
*his arms*  
*man bites*  
*white dog*

Grammar  
in → *his arms*  
a → *man*  
a → *white dog*

Output Sentence  
*A man bites*  
*a white dog*  
*in his arms.*

文の一部で重要そうなものを複数推定

↑  
“キーフレーズ”

[Ushiku+, ACM MM 2012]では：  
Fisher Vector + 線形分類オンライン学習

CVPR 2015 の各論文では：  
CNN（オンライン学習なのは一緒）

文法モデルを利用して繋ぎ、説明文に

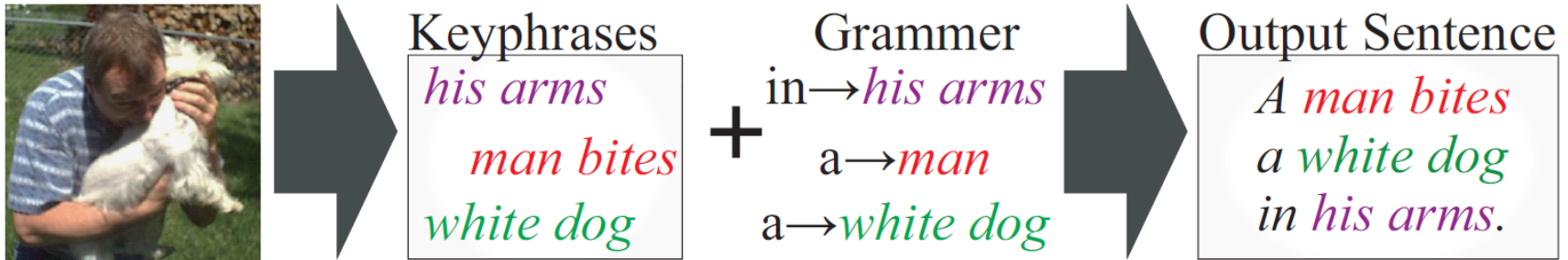
[Ushiku+, ACM MM 2012]では：  
キーフレーズと文法モデル、  
ビームサーチで文をつなぐ

CVPR 2015 の各論文では：  
RNNとビームサーチで文をつなぐ

- いずれも画像+キャプションのみから学習可能
- 全体の流れは非常に似ている

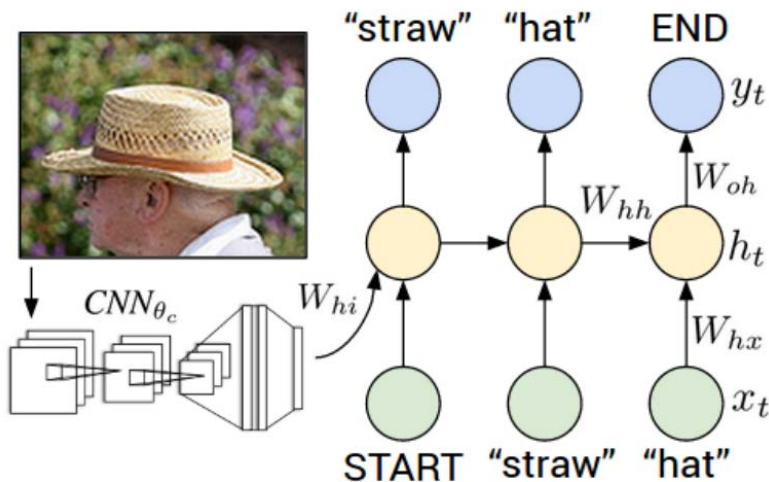
# 一番大きく違うところは...?

- 深層学習以前の新規キャプション生成



何らかの語句に変換してから文生成器へ

- 深層学習による新規キャプション生成

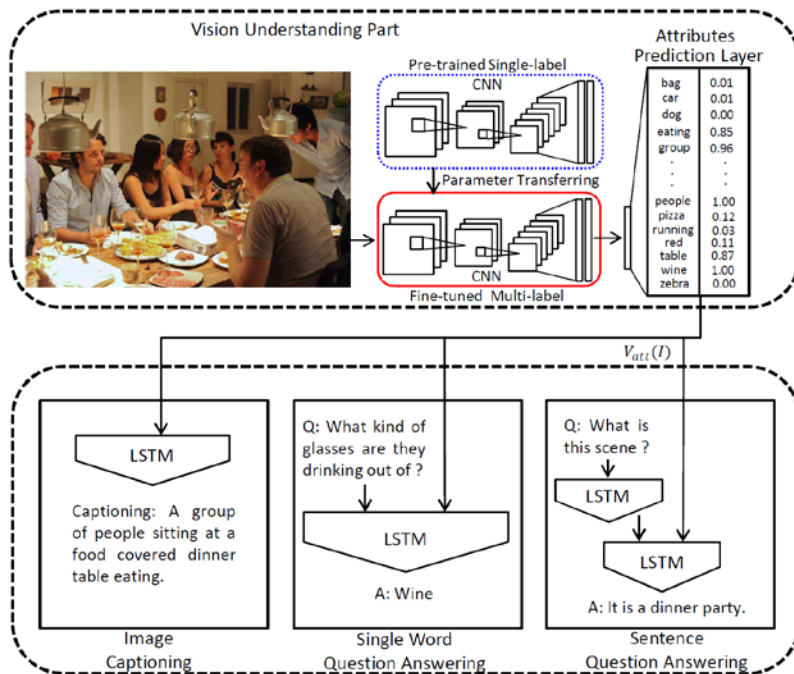


画像特徴量を直接文生成器へ

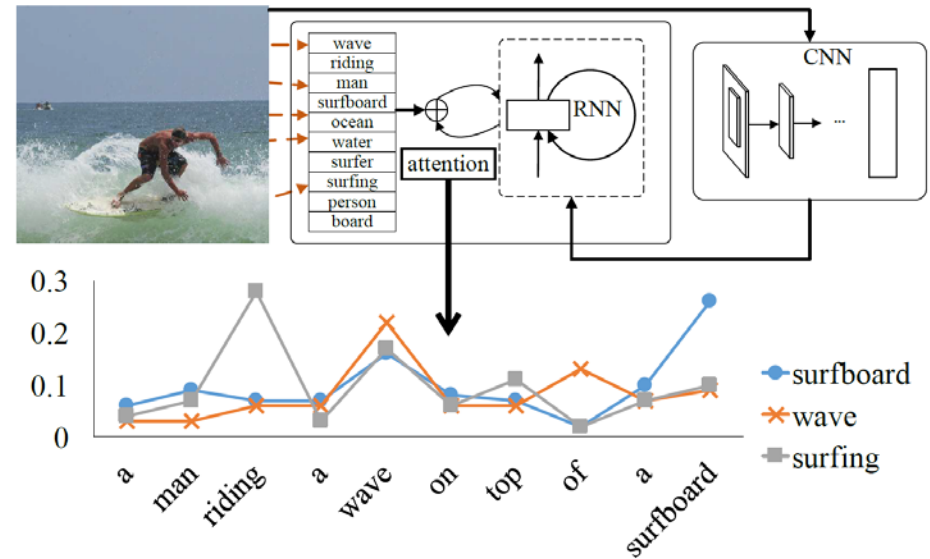
# ところが最近では...

- CNNで事物の認識まで済ませてRNNで文生成[Wu+, CVPR 2016][You+, CVPR 2016]

→ **画像特徴量の段階でRNNに渡すより高性能！**



[Wu+, CVPR 2016]



[You+, CVPR 2016]

- **深層学習以前のアプローチとより類似**



# 画像認識分野とNLPとの融合

- 2分野が融合して新たに生まれたものの例：
  - アテンションモデルの利用 [Xu+, ICML 2015]



A woman is throwing a frisbee in a park.



A dog is standing on a hardwood floor.

- 画像+キャプションから注視モデルも学習！



A



bird



flying



over



a



body



of



water



.

# NLPと画像理解の発展

# 現在の展開：精度の発展

- 画像認識

InceptionモデルやResNetなど、より高精度なCNN

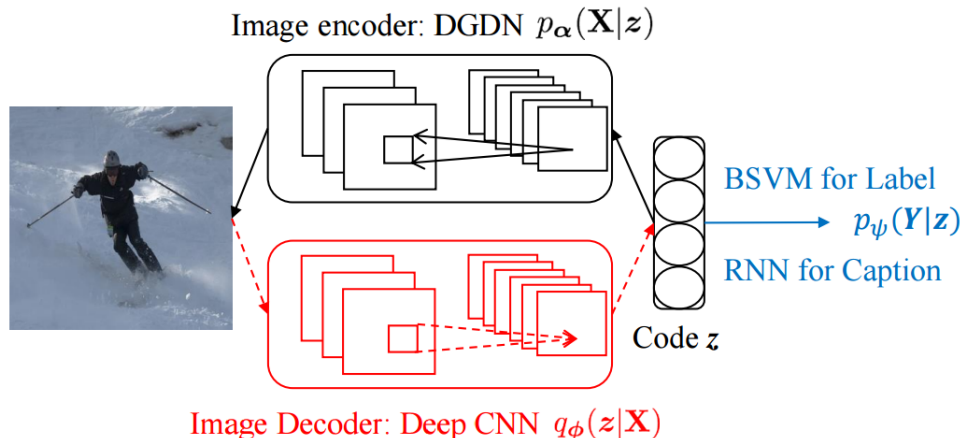
- 自然言語処理

画像認識側が完璧になったと仮定した文生成

[Gupta+Mannem, ICONIP 2012][Elliott+Keller, EMNLP 2013][Yatskar+, \*Sem 2014][Yao+, ICLR workshop 2016]

- 機械学習

変分自己符号化器の利用 [Pu+, NIPS 2017]

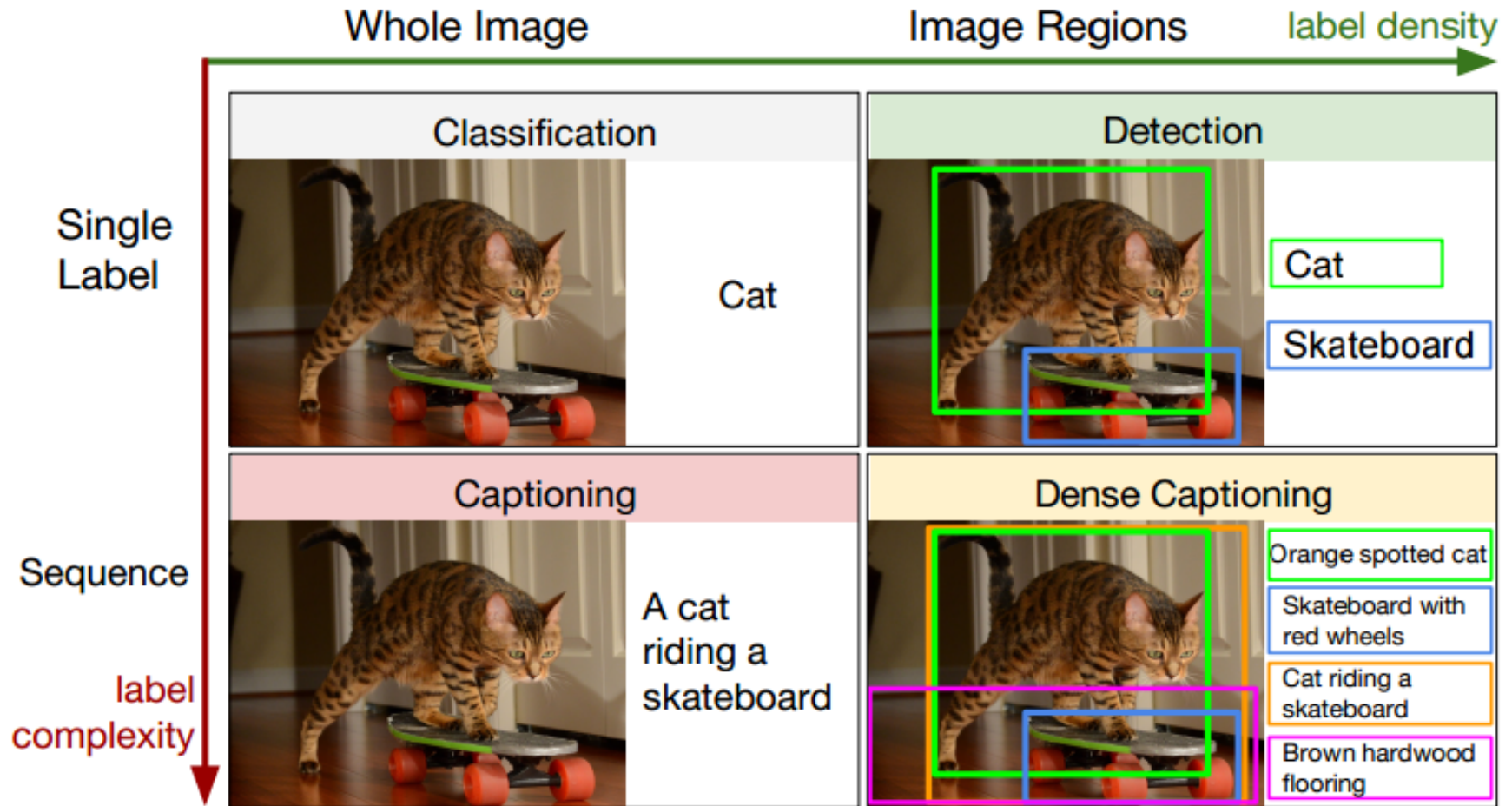




# 現在の展開：問題の発展

## より細かいキャプション生成

[Lin+, BMVC 2015] [Johnson+, CVPR 2016]



# 現在の展開：問題の発展

## アルバムのような系列画像にキャプション生成

[Park+Kim, NIPS 2015][Huang+, NAACL 2016]



The family got together for a cookout.



**They** had a lot of delicious food.



The dog was happy to be **there**.



**They** had a great time on the beach.



**They** even had a swim in the water.

# 現在の展開：問題の発展

## 感性語Sentiment Termを重視したキャプション生成

[Mathews+, AAAI 2016][Andrew+, BMVC 2016]←Ours!

ニュートラルな文



This is a dog resting on a computer.  
A white shaggy beautiful dog laying its  
head on top of a computer keyboard.

ポジティブな文  
(生成した例)

A motorcycle parked behind a truck  
on a green field.  
A beat up, rusty motorcycle on  
unmowed grass by a truck and trailer.



# 動画キャプション生成

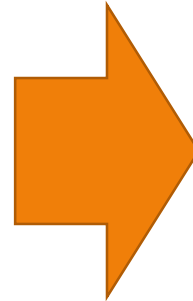
[Andrew+, ICIP 2016]



A man is holding a box of doughnuts.  
Then he and a woman are standing next each other.  
Then she is holding a plate of food.

# 他言語化・キャプション翻訳

[Hitschler+, ACL 2016]



A pole with two lights  
for drivers. (英語)

Ein Masten mit zwei Ampeln  
für Autofahrer. (独語)



# キャプションからの画像生成

[Zhang+, 2016]

This bird is blue with white and has a very short beak.

(この鳥は白の入った青色で、とても短いくちばしをもっています。)



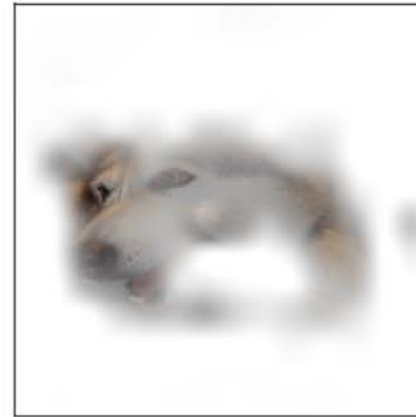
This flower is white and yellow in color, with petals that are wavy and smooth.

(この花は白と黄色で、波打った滑らかな花びらをもっています。)



# ビジュアル質問応答

[Fukui+, EMNLP 2016]



What vegetable is the dog  
chewing on?

MCB: carrot

GT: carrot

What kind of dog is this?

MCB: husky

GT: husky

What kind of flooring does  
the room have?

MCB: carpet

GT: carpet



What color is the traffic  
light?

MCB: green

GT: green

Is this an urban area?

MCB: yes

GT: yes

Where are the buildings?

MCB: in background

GT: on left

まとめ



# 第11回の内容と目標

## 1. 座学 NLPと画像理解

- LSTMとword2vec (SGNS)の復習
- ニューラル機械翻訳とアテンション
- 画像キャプション生成

## 2. 演習 アテンションを用いた機械翻訳モデルの実装

## 3. 演習 アテンションを用いたキャプション生成モデルの実装

## 4. 座学 NLPと画像理解の発展

- 画像キャプション生成やその他の課題へ

# 第9回～第11回の内容と目標

- **基礎：**  
系列データの理解/生成に用いられる  
Recurrent Neural Networks (RNNs) の理解
- **応用：**  
自然言語処理や画像 + 言語融合分野の理解

Chap. 9 RNN

Chap. 10 Long Short-Term Memory (LSTM)  
自然言語処理 (NLP)

Chap. 11 NLPと画像理解